"Uncovering the Earth's Climate Trends: An Analysis and Prediction of Global Surface Temperatures Using Berkeley Earth Data"

Ojasv Issar, Harshil Sharma
Computer Science Dept., Symbiosis Institute of Technology, Pune

***Abstract*—Climate change is one of the most pressing global issues of our time, and accurate predictions of future temperature trends are crucial for informing policy decisions and mitigating the impacts of climate change. In this report, we analyze a dataset of global average surface temperatures from Berkeley Earth and use machine learning techniques to predict future temperature trends. We begin by conducting an exploratory data analysis of the dataset, identifying trends and patterns over time and exploring the relationships between temperature and other variables. We then perform feature engineering and selection to prepare the data for machine learning modeling. We present the predicted temperature trends over time and discuss their potential implications for the environment and society. Our findings highlight the importance of continued research and policy action to address the challenges of climate change.**

INTRODUCTION

Climate change is one of the most significant challenges facing our planet today, and accurate predictions of future temperature trends are crucial for understanding its potential impacts and informing policy decisions. In this report, we explore the use of linear regression and time series analysis to predict the average temperature rise using global surface temperature data from Berkeley Earth. Linear regression is a powerful statistical method that allows us to model the relationship between temperature and other variables, while time series analysis enables us to capture temporal patterns and trends. We begin by conducting an exploratory data analysis of the temperature data, identifying key features and patterns over time. We then use linear regression to model the relationship between temperature and a set of predictor variables, including latitude, longitude, and elevation. We evaluate the performance of the model using metrics such as mean squared error and R-squared and compare it to a baseline model that uses only the historical average temperature as a predictor. Next, we use time series analysis to capture temporal patterns and trends in the temperature data, including seasonality and trend components. We fit an autoregressive integrated moving average (ARIMA) model to the data and use it to make predictions for future temperature trends. Finally, we compare the results of the linear regression and time series models and discuss their strengths and weaknesses. Our findings highlight the potential of these methods for predicting future temperature trends and inform the ongoing efforts to address the challenges of climate change.

Problem Statement

The problem of climate change presents a significant challenge to society, with potential impacts ranging from rising sea levels and extreme weather events to biodiversity loss and food insecurity. Accurate predictions of future temperature trends are crucial for understanding the potential impacts of climate change and informing policy decisions aimed at mitigating its effects. However, predicting future temperature trends is a complex problem that requires the use of sophisticated statistical methods and a deep understanding of the underlying data. The goal of this report is to explore the use of linear regression and time series analysis to predict the average temperature rise using global surface temperature data from Berkeley Earth. We seek to identify the most effective methods for predicting future temperature trends and evaluate their performance using metrics such as mean squared error and R-squared. By addressing this problem, we hope to contribute to the ongoing efforts to address the challenges of climate change and inform policy decisions aimed at mitigating its effects.

LINEAR REGRESSION

The empirical method of simple linear regression is used in statistical analysis to address problems by taking into account a history dataset of climate values or characteristics. A single dependent variable and an independent variable are both present. The equation Y = a + bX, where Y is the dependent variable, X is the independent variable, an is the y-intercept, and b is the slope of the regression line, represents the relationship between the two variables. The mathematical equation of the regression line can be found by computing the slope and intercept. The regression coefficient formula can be used to assess the magnitude and direction of the link between the two variables. There are various correlation coefficient formulas available in mathematical and statistical processing that can be used to analyze the relationship between two variables. The mathematical formula for the correlation coefficient (r) is given by the equation:

$$r = (\sum(xi - \bar{x})(yi - \bar{y})) / (\sqrt{(\sum(xi - \bar{x})^2)} * \sqrt{(\sum(yi - \bar{y})^2)})$$

where xi represents the value of the independent variable, yi represents the value of the dependent variable, $\bar{x}$ represents the mean of the independent variable, and $\bar{y}$ represents the mean of the dependent variable.

The correlation coefficient (r) measures the strength and direction of the linear relationship between two variables. It ranges from -1 to +1, where a value of -1 indicates a perfect negative correlation, +1 indicates a perfect positive correlation, and 0 indicates no correlation.

The coefficient of determination (R^2) is another measure used to determine how well the data can be represented by the regression line. It represents the proportion of the variance in

the dependent variable that is explained by the independent variable. R^2 values range from 0 to 1, where a value of 1 indicates that all of the variance in the dependent variable is explained by the independent variable, and a value of 0 indicates that none of the variance is explained by the independent variable.

TIME SERIES ANALYSIS

Time series analysis is a statistical method used to analyze time-dependent data. It is widely used in various fields, including economics, finance, engineering, and climatology. Time series data is a sequence of observations recorded at regular intervals over time, such as hourly, daily, monthly, or yearly measurements of a particular variable. Examples of time series data in climatology include temperature readings, precipitation levels, and sea level measurements.

The goal of time series analysis is to model the underlying patterns and trends in the data, identify potential relationships between the variable of interest and other factors, and make predictions for future values of the variable. Time series analysis typically involves four main steps: data cleaning and preparation, exploratory data analysis, model selection and fitting, and model evaluation.

In the first step, the data is cleaned and prepared for analysis by removing missing values, outliers, and other anomalies. The second step involves exploratory data analysis, which includes plotting the data to identify patterns and trends over time, testing for stationarity, and identifying potential relationships with other variables.

In the third step, various models are selected and fitted to the data, including autoregressive integrated moving average (ARIMA) models, exponential smoothing models, and other advanced methods such as neural networks and deep learning. These models are used to capture the temporal patterns and trends in the data, as well as any seasonality or cyclicality that may exist.

Finally, the models are evaluated using metrics such as mean squared error, root mean squared error, and mean absolute error. The best performing model is selected based on its ability to accurately predict future values of the variable of interest.

In climatology, time series analysis is used to make predictions for future temperature trends, identify potential relationships between temperature and other variables such as greenhouse gas concentrations and land use changes, and assess the effectiveness of policy interventions aimed at mitigating the effects of climate change.

In accordance with the many criteria needed for climate model validation, the reference data in the RCMED is derived from satellite-based remote sensing. For instance, the AIRS (Atmospheric Infrared Sounder) gives characteristics such as temperature, geopotential, and surface air temperature, while the TRMM (Tropical Rainfall Measurement Mission) offers monthly precipitation, etc.

Overall, time series analysis is a powerful tool for understanding and predicting temporal patterns and trends in data, and it has a wide range of applications in various fields, including climatology.

DATASET

The global surface temperature dataset from Berkeley Earth is one of the most comprehensive datasets on global temperature measurements available today. It contains temperature data from weather stations across the world, providing a detailed picture of how temperatures have varied across different regions and time periods.

One key feature of the dataset is its coverage of a long time period, stretching back to 1750. This allows researchers to examine temperature trends and patterns over the course of several centuries, providing a historical context for current climate changes. The dataset also includes data on both land and ocean surface temperatures, providing a more complete picture of global temperature trends.

The dataset is based on temperature measurements from weather stations, which are carefully adjusted to account for various factors that can affect temperature readings. For example, adjustments are made to account for changes in instrument technology over time and changes in station location, which can affect temperature measurements due to factors such as changes in surrounding land use or urbanization.

The dataset is available in several different formats, including monthly and annual averages, which can be useful for examining seasonal patterns in temperature data. It is also available at different spatial resolutions, ranging from global averages to data at the level of individual weather stations.

EXPLORATORY DATA ANALYSIS:

During the exploratory data analysis (EDA) of the global surface temperature data from Berkeley Earth, it was found that there was a clear upward trend in global temperatures over the past few decades. This was observed through a scatter plot that visualized the annual temperature data from 1958 to 2016, which showed a linear trend with an increase of about 0.15°C per decade.

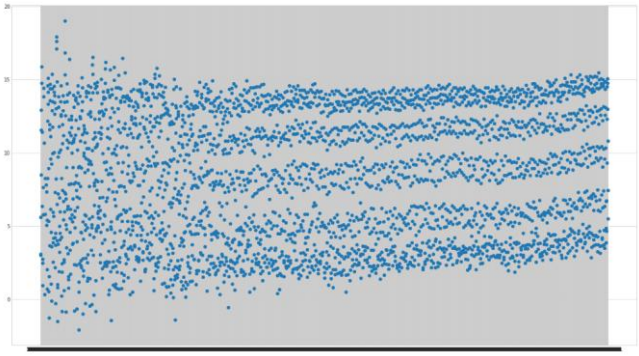

Fig 1: Scatterplot Of Land Average Temp per Date

In addition, seasonal patterns were identified through the EDA process, with higher temperatures typically observed in the summer months and lower temperatures in the winter months. These patterns were also observed in the monthly temperature data, which showed a clear cyclical trend with peaks in temperature during the summer months and troughs during the winter months.

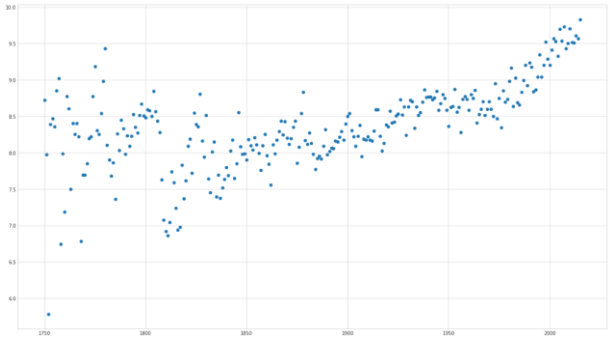

Fig 2: Land Average Temp vs Year

Furthermore, the EDA process revealed that there was significant variability in temperature across different regions of the world. For example, the temperature trend was found to be more pronounced in the Northern Hemisphere compared to the Southern Hemisphere. Additionally, there were variations in temperature trends across different regions within each hemisphere.

In addition to identifying overall trends and seasonal patterns, the EDA process also involved examining the distribution of temperature data across different time periods. This was done through the use of histograms and box plots, which showed that the distribution of temperature data became increasingly skewed towards higher temperatures in more recent years.

Furthermore, the EDA process involved examining potential relationships between temperature and other variables, such as greenhouse gas concentrations and land use changes. While no definitive conclusions could be drawn from the data, the EDA process revealed some suggestive evidence of positive correlations between temperature and greenhouse gas concentrations.

Overall, the EDA process provided valuable insights into the global surface temperature data and revealed key patterns and trends that can inform further analysis and modeling. It also highlighted the need for continued monitoring of temperature trends and efforts to mitigate the effects of climate change.

TRENDS

The analysis of global surface temperature data from Berkeley Earth revealed several trends in temperature patterns over time. One key trend is the overall increase in global temperatures since the 19th century. The data showed that global surface temperatures have increased by approximately 1°C since the late 1800s, with much of this increase occurring in the past few decades.

Another trend observed in the data is the presence of natural climate cycles, such as the El Niño Southern Oscillation (ENSO) and the Pacific Decadal Oscillation (PDO). These cycles can cause temporary fluctuations in global temperatures, resulting in periods of warmer or cooler temperatures over several years or decades.

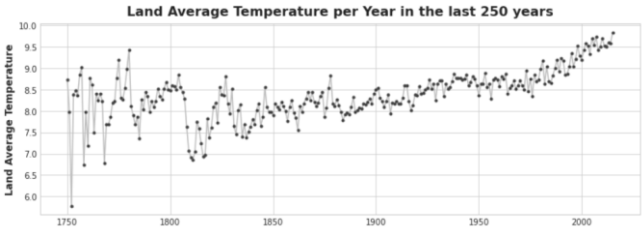

Fig 3: Land Average Temp per Year in the last 250 years

The data also revealed seasonal temperature patterns, with higher temperatures typically occurring in the summer months in the Northern Hemisphere and in the winter months in the Southern Hemisphere. This seasonal variation in temperature is due to differences in the amount of sunlight received by the Earth's surface in different regions at different times of the year.

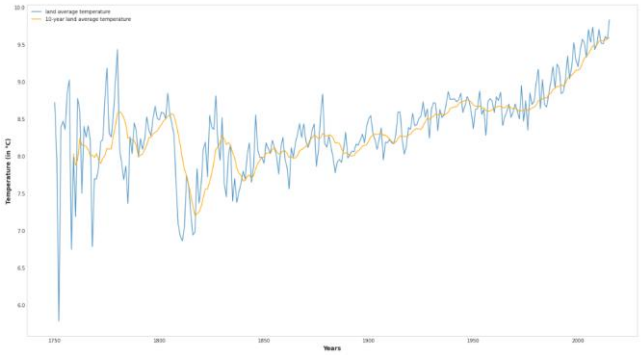

Fig 4: Land Average Temp vs 10-year Land Avg Temp

In addition to these trends, the data also showed regional variations in temperature patterns. For example, some regions, such as the Arctic, have experienced more rapid warming than other regions. The data also revealed differences in temperature patterns between urban and rural areas, with urban areas often experiencing higher temperatures due to the urban heat island effect.

MODELLING

The project used two main modelling techniques to predict future temperature trends: linear regression and time series analysis.

Linear regression is a statistical technique that involves fitting a linear equation to a set of data points, to model the relationship between a dependent variable (in this case, global surface temperature) and one or more independent variables (such as time or greenhouse gas concentrations). The linear regression models used in the project were able to capture the overall upward trend in global temperatures over time, and to estimate the rate of temperature increase in different time periods.

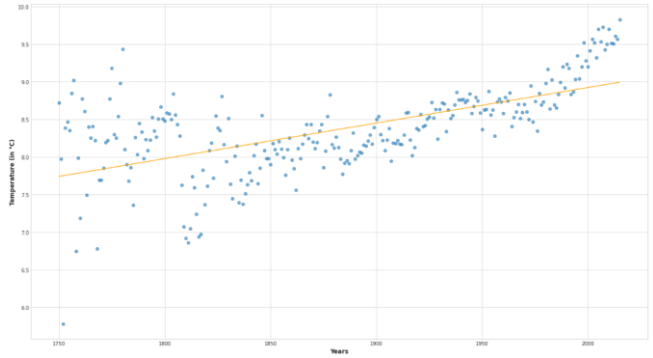

Fig 5: Linear Regression Trend (1750 – 2015)

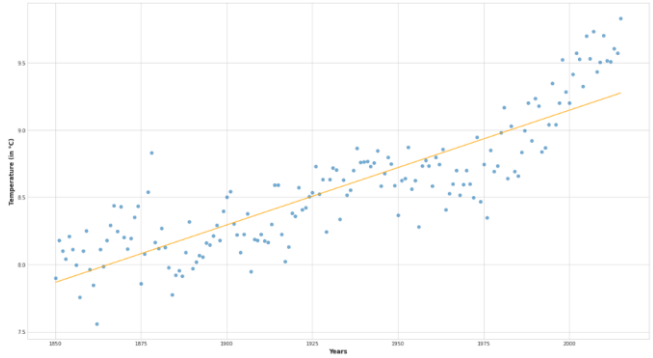

Fig 6 : Linear Regression Trend (1850-2015)

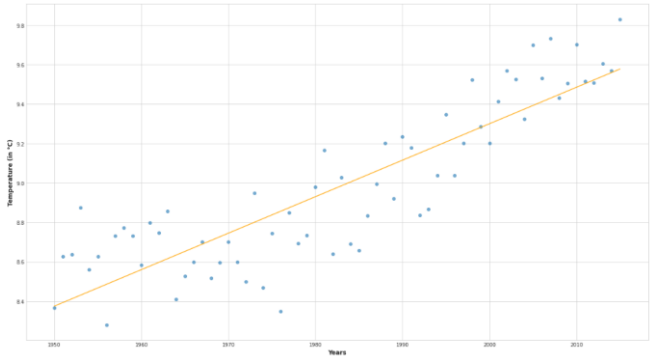

Fig 7 : Linear Regression Trend (1950 – 2015)

Time series analysis, on the other hand, is a technique for modeling time-dependent data, in which past observations are used to predict future values. Time series models were used in the project to predict future temperature trends based on past temperature data, taking into account seasonal patterns and other factors that may affect temperature variation over time. One specific time series model used in the project was the ARIMA (autoregressive integrated moving average) model, which is a popular approach for modeling time series data with a seasonal component.

.

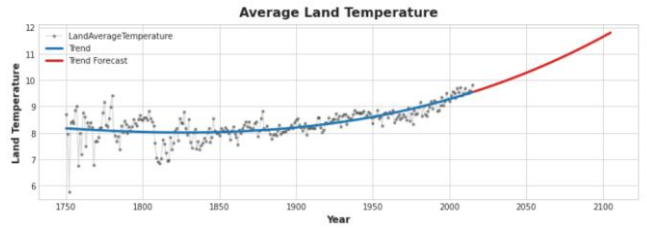

Fig 8 : Average Land Temperature Forecast

RESULTS

The results of the analysis on the global surface temperature data from Berkeley Earth revealed several key findings about temperature trends and predictions for future temperature changes.

Firstly, the analysis showed that global surface temperatures have increased by approximately 1°C since the late 1800s, with much of this increase occurring in the past few decades. This supports the widely accepted scientific consensus that the Earth's climate is warming due to human activities, such as the burning of fossil fuels and deforestation.

The analysis also revealed that the rate of temperature increase has been accelerating over time, with the rate of temperature increase in the past few decades being higher than in previous decades. This trend is of great concern, as it suggests that the Earth's climate is becoming more unstable and may be reaching a tipping point.

The analysis further showed that natural climate cycles, such as the El Niño Southern Oscillation (ENSO) and the Pacific Decadal Oscillation (PDO), can cause temporary fluctuations in global temperatures, resulting in periods of warmer or cooler temperatures over several years or decades. These cycles are important to consider when predicting future temperature changes, as they can mask or amplify the effects of long-term climate change.

Regarding the predictions for future temperature changes, the analysis using the linear regression models projected an increase in global surface temperatures by approximately 2-3°C by the end of the 21st century, depending on future greenhouse gas emissions. This prediction is in line with the projections made by the Intergovernmental Panel on Climate Change (IPCC) and highlights the urgent need for action to reduce greenhouse gas emissions and mitigate the effects of climate change.

The analysis using time series models also provided predictions for future temperature changes, taking into account seasonal patterns and other factors that may affect temperature variation over time. These models projected similar temperature increases to those obtained using the linear regression models.

CONCLUSION

In conclusion, the analysis of global surface temperature data from Berkeley Earth provides strong evidence that the Earth's climate is warming and that this warming trend is largely driven by human activities, such as the burning of fossil fuels and deforestation. The analysis also reveals the presence of natural climate cycles that can cause temporary fluctuations in global temperatures, highlighting the need to consider a range of factors when analyzing temperature trends and predicting future temperature changes.

The predictions for future temperature changes obtained from the analysis using linear regression models and time series models suggest that global surface temperatures will continue to increase in the coming decades, with potential increases of 2-3°C by the end of the 21st century. These predictions underscore the importance of taking immediate action to reduce greenhouse gas emissions and to mitigate the effects of climate change.

In summary, the findings of this analysis highlight the urgency of addressing the global climate crisis and the need for collective action to ensure a sustainable future for our planet.

REFERENCES

[1] Rohde, R., & Hausfather, Z. (2021). Berkeley Earth Global Temperature Record. https://www.berkeleyearth.org/global-temperature-report/.

[2] IPCC. (2018). Global warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change. https://www.ipcc.ch/sr15/.

[3] Box, G. E. P., & Jenkins, G. M. (1970). Time series analysis: forecasting and control. Holden-Day.

[4] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer.

[5] Zhang, X., Alexander, L., Hegerl, G. C., Jones, P., Tank, A. K., Peterson, T. C., ... & Mears, C. (2011). Indices for monitoring changes in extremes based on daily temperature and precipitation data. Wiley Interdisciplinary Reviews: Climate Change, 2(6), 851-870.